Analysis of teak (*Tectona grandis*) genome and its diversity

Songkran Thongthawee[1, 2] and Hugo Volkaert [1, 2, 3,*]

[1]*Center for Agricultural Biotechnology, Kasetsart University, Kamphaeng Saen Campus, Nakhon Pathom 73140, Thailand*
[2]*Center for Agricultural Biotechnology: (AG-BIO/PERDO-CHE), Bangkok 10900, Thailand*
[3]*Plant Research Group, National Center for Genetic Engineering and Biotechnology (BIOTEC), Thailand Science Park, Klong Luang, Pathum Thani, 12120*
**e-mail: kpshgv@ku.ac.th*

Abstract

The genomes of ten teak trees were sequenced using Illumina Technology. One tree was sequenced with 5 fragment libraries (200, 300, 500, 5,000 and 10,000 bp), while the other 9 were sequenced with only two short insert libraries. The complete chloroplast genome was assembled, but only fragments of the mitochondrial genome have been obtained. A set of chromosomal contigs and scaffolds was assembled from the reference tree. The longest contig was more than 600,000 bp and the longest scaffold was 2.6 Mbp. The total sequence length is approximately 300 Mbp, which accounts for probably more than 80% of the actual chromosomal genome. The scaffolds cannot yet be assembled into chromosomal maps as a good mapping population is not available. Gene prediction software indicates the presence of more than 36,000 protein coding genes. Gene models have been developed for a set of general metabolic enzymes. Three different chloroplast haplotypes were detected among the 10 trees, differing from each other by one SNP and one 10 bp insertion/deletion. SNP detection among the chromosomal genomes of the 10 trees found 643,647 nucleotide substitutions, which amounts to 1 SNP per 500 bp.

Keywords: Illumina Technology, chloroplast genome, mitochondrial genome, chromosomal genome, SNP

Introduction

Teak (*Tectona grandis* Linn. fil.) is an important forest tree of large size in the deciduous forest ecosystems of India, Myanmar, Thailand, Laos and Java. It produces a timber highly valued for its durability and workability, being resistant to fungal rot and termite attack (Lukmandaru and Takahashi 2008). It is popular for use as structural timber or for decorative purposes both interior and exterior. Teak is a fast growing tree species that has been established in plantations both within and outside its natural area of origin. Thailand was a major exporter of teak logs during the first half of the twentieth century but supplies dwindled due to overharvesting and illegal logging. Conversion of forest areas for agriculture and degradation of the remaining forests further threaten forest tree genetic resources, including teak. Currently, the Thai government emphasizes forest restoration and production of timber from plantation forests (FAO 2001). Knowledge about the remaining genetic diversity is urgently needed for teak and many other forest tree species such that it can be better conserved and utilized through tree improvement programmes. Teak has a diploid genome with 18 pairs of chromosomes. Little is known though about the genetic diversity of

teak. To develop a database of molecular genetic polymorphisms in teak, the genome of ten trees has been sequenced using the Illumina Technology.

Methodology

Nine trees from the Royal Forest Department's clonal teak germplasm conservation bank in KhonKaen and Phitsanuloke and one tree growing on the Kasetsart University Kamphaengsaen campus, were used for this study. Nuclei were isolated in order to avoid as much as possible sequencing the chloroplast and mitochondrial genome. Forty grams of young fresh leaf were homogenized in cold IS1 buffer (1X buffer IS1: 0.022 M EDTA, 0.02 M Tris base, 0.16 M KCl, 1.5% sucrose) and filtered through 4 layers cheesecloth and one layer of miracloth (Calbiochem). The filtrate was centrifuged at 2,000 g, 4°C, 20 min and the green supernatant was discarded. The pellet was resuspended in IS1 buffer with 2-mercaptoethanol 400 µl and 1 ml of tritonX-100 and incubated on ice for 30 min. The solution was centrifuged again at 2,000 g, 4°C for 20 min and the pellet resuspended in 1 ml IS1 buffer. The solution was carefully layered on top of a step gradient of 50%, 30% and 20% of sucrose solution in 15-ml centrifuge tube. After centrifugation at 2,000 g, 4°C, 20 min in a swinging bucket rotor the sucrose solution was carefully removed and the pellet resuspended in 1 ml IS1 buffer in a new 1.5 ml microcentrifuge tube. The nuclei were harvested by another centrifugation for 5 min at 12,000-13,000 rpm, room temperature and nuclear DNA was extracted from the pellet by Genomic DNA Mini Kit (Plant), (Geneaid, Taiwan) following the manufacturer's protocol.

The genomic DNA was fragmented and libraries were constructed for each tree with insert sizes of 200 bp and 300 bp. For one tree additional libraries were constructed with insert sizes 500, 5,000 and 10,000 bp. The libraries were sequenced using paired-end reads of 101 bp with the Illumina HiSeq2000.

For the chloroplast assembly the raw sequences were searched by BLASTN (Camacho et al. 2009) with tomato chloroplast as a query. For those reads having high similarity to the tomato chloroplast (e-value <0.0001) the matching pairs were extracted from the dataset. Assembly of the chloroplast genome using the selected paired reads was achieved by Edena (Hernandez et al. 2014) and finishing manually. A similar approach was followed for the mitochondrial genes, though no attempt was made to obtain a complete mitochondrial genome. Chromosomal genome assembly was attempted by SOAPdenovo2 (Luo et al. 2012), MASURCA (Zimin et al. 2013), ALLPATHS (MacCallum et al. 2009), dipSPAdes (Safonova et al. 2014) and Platanus (Kajitani et al. 2014).

Protein coding genes have been detected and annotated by AUGUSTUS (Stanke et al. 2008). SNP discovery has been done by aligning of reads to the preliminary reference genome with bwa (Li, 2013) and filtering for SNPs with samtools and bcftools. Alternatively SNPs were detected from the unaligned sequences by discoSnp (Quillery et al. 2014).

Results

For each 200 or 300 bp fragment sequencing library, between 37,190,262 and 45,187,349 paired reads were obtained which amounts to 20 to 25 fold coverage of a 360 Mbp genome. The assembly of the chloroplast genome by Edena and manual finishing resulted in a circular genome of 153,943 or 153,953 bp. Three haplotypes were detected among the 10 trees, differing from each other by a 10-bp insertion/deletion and one single nucleotide substitution.

The sequences have been deposited in DNA database repositories with accession numbers HF567869-HF567871.

The mitochondrial genome has been partly reconstructed from Edena contigs. The sequence coverage of the mitochondrial genome was about 8-10 x higher than the chromosomal sequences, while the chloroplast genome was sequenced again 10 x higher than the mitochondrial genome. Based on the coverage information, 56 Edena contigs amounting to approximately 400 kbp were selected and compared to mitochondrial genomes of other plant species. These contigs were then assembled into 16 larger contigs.

The chromosomal genome has been assembled by SOAPdenovo2, MASURCA, ALLPATHS, dipSPAdes and Platanus. For all approaches the total genome length ranged from 160 to 380 Mbp. The Platanus assembly resulted in the longest scaffolds up to 2.6 Mbp. The dipSPAdes assembly contained the longest contigs, with the longest one 647,993 bp.

For the Platanus assembly, 15,846 scaffolds of 1000 bp or longer were obtained, summing up to 289,899,029 bp, including 24,522,536 Ns (8.46%) The 17,416 dipSPAdes haplocontigs longer than 1000 bp summed up to a total of 225,176,024 bp which did not contain a single N. The GC content of the assemblies is about 32%.

Table 1 Summary of Illumina sequencing results per library.

Libraries	Sample	Total Bases	No. of fragment	N (%)	GC (%)	Q20 (%)	Q30 (%)
200 bp	S1	9,127,844,498	45,187,349	0.011	36.02	94.96	88.97
	S2	8,471,719,410	41,939,205	0.011	35.87	94.98	89.06
	S3	8,805,227,470	43,590,235	0.011	36.13	94.81	88.83
	S4	9,060,299,536	44,852,968	0.011	36.36	94.95	88.91
	S5	7,973,922,326	39,474,863	0.011	36.31	94.63	88.59
	S6	8,756,864,832	43,350,816	0.011	35.66	95.04	89.24
	S7	8,932,722,194	44,221,397	0.010	36.14	95.74	90.68
	S8	8,751,553,040	43,324,520	0.010	36.42	95.63	90.44
	S9	9,064,508,610	44,873,805	0.010	35.64	95.49	90.39
	S10	8,447,043,898	41,817,049	0.010	35.98	95.83	90.57
300 bp	S1	8,037,612,724	39,790,162	0.045	33.01	92.99	85.62
	S2	8,799,241,806	43,560,603	0.043	32.97	93.25	85.91
	S3	8,348,521,024	41,329,312	0.043	33.19	93.50	86.36
	S4	8,117,393,230	40,185,115	0.044	33.14	93.70	86.68
	S5	8,081,425,514	40,007,057	0.044	33.31	93.29	86.01
	S6	8,007,254,952	39,639,876	0.034	33.18	94.39	88.03
	S7	8,631,501,006	42,730,203	0.035	33.28	94.18	87.71
	S8	7,622,322,944	37,734,272	0.033	33.68	93.71	86.88
	S9	7,512,432,924	37,190,262	0.033	33.91	93.79	87.04
	S10	7,695,880,638	38,098,419	0.035	33.55	94.17	87.72
500 bp	S1	7,543,668,992	37,344,896	0.057	33.41	93.89	87.18
	S2	7,291,245,752	36,095,276	0.056	32.68	93.77	86.98
5 kb	S1	6,563,160,184	32,490,892	0.116	36.53	91.17	86.07
10 kb	S1	10,616,764,480	52,558,240	0.004	36.18	90.16	84.48

Using the dipSPAdes contigs, gene discovery and annotation by AUGUSTUS with tomato as a reference resulted in 36,206 protein coding entries. For some protein coding genes (some general metabolism genes, nitrogen uptake and assembly genes, some hormone receptors and response genes) manual curation has been done which mostly confirmed the AUGUSTUS suggestions, though in other cases adjustments were in order.

SNP discovery by aligning of reads to without a reference genome as done by discoSnp detected a total of 643,647 SNPs of which 389,741 were transitions and 253,906 were transversions. Least common among the SNPs were C/G transversions.

Table 2: Summary of SNPs found among the 10 teak trees.

	Transition		Transversion				Overall
Substitution	**C / T**	**A / G**	**G / T**	**A / C**	**A / T**	**C / G**	
No. of SNPs	193,152	196,589	61,649	63,254	77,104	51,899	643,647
Total of SNPs	389,741		253,906				

Discussion

Teak is an important tree species in the monsoon deciduous forest ecosystems of India, Myanmar, Thailand and Java. It is also widely planted in other tropical regions because of the highly valued timber and fast growth rate. However, very little is known about the genetic diversity of this species. In order to get insight into the genetic diversity the genome of 10 teak trees from Thailand was sequenced using the Illumina HiSeq2000 approach. Even though only 101 bp DNA sequence reads are obtained by this technology, and it is difficult to obtain good quality sequence data from large fragments, a large proportion of the genome sequence has become accessible. The teak nuclear DNA content was estimated to be slightly lower than rice (C-value database, Bennett and Leitch, 2012), though flow cytometry analysis indicates that it might be slightly higher. Therefore, we estimate the teak genome length to be 360 Mbp. The dipSPAdes contigs sum up to 225 Mbp, or about 62% of the total genome, while the Platanus scaffolds (290 Mbp) amount to 80% of the genome.

Though the analysis of the teak genome is still in an initial phase, some conclusions can be drawn already. Both the Platanus programme and the dipSPAdes module in the SPAdes programme are “diploid-aware” genome assemblers, while SOAPdenovo, MASURCA and ALLPATHS-LG have not been designed to handle sequence data obtained from heterozygous genomes. The Platanus programme uses a relatively greedy algorithm to construct the scaffolds, while dipSPAdes is rather conservative. The dipSPAdes contigs might still be assembled into larger scaffolds with programme like BESST (Sahlin et al., 2014).

Since a linkage mapping population for teak has not been developed, the contigs and scaffolds cannot be further assigned to the individual chromosomes. However, the developed resources will be very useful when such a mapping population becomes available. However, it is clear that even without a complete reference genome, the genetic diversity within the Thai teak populations can be studied. The contigs and scaffolds can be used as references for the alignment of the individual reads (bwa-mem) from the 10 trees and SNPs can be detected using various tools (samtools, bcftools, GATK etc), or alternatively SNP can be discovered without a reference genome using discoSnp. More than 500,000 SNPs were discovered by both bwa-mem and discoSnp. These SNPs still need to be validated and cross-checked, but preliminary results indicate that the SNPs detected by discoSnp are most likely reliable. For those SNPs that were checked by further in silico analysis, all polymorphisms could be recovered. A collection of 500,000 SNPs among Thai teak would amount to one SNP per 700 bp. Thai teak is far less polymorphic than teak from Southern India. Thus over its total geographical range, the number of polymorphisms in the teak genome can be expected to be much higher. In addition to the SNPs, several 1000 microsatellite loci have been identified. However, most of the microsatellite sequences were at breaks between contigs.

Gene annotation with the AUGUSTUS programme indicates a high number of gene models, several of which though are quite short and may not be true protein coding genes. Most of the longer protein coding sequences were highly similar to known protein coding sequences form other plant species using BLASTn with the NR protein database or with the EST nucleotide database. Most similar hits were from related species such as *Mimulus gutatus* but also tomato and cocoa genes were frequently found among the top hits.

Conclusion

Forest tree breeders have long lamented about the lack of molecular genetic resources for plantation species. Here we attempt to build a molecular genetic resource for teak diversity and breeding by giving access to most of the protein coding sequences in the teak genome and its diversity among a set of teak trees from Thailand. These resources will be made publicly available through a website as a searchable database of gene and protein sequences and SNPs.

Acknowledgements

This research is supported by the Center of Excellence on Agricultural Biotechnology, Science and Technology Postgraduate Education and Research Development Office, Commission on Higher Education, Ministry of Education. (AG-BIO/PERDO-CHE).

References

Bennett M.D., Leitch I.J. (2012) Plant DNA C-values database (release 6.0, Dec. 2012) http://www.kew.org/cvalues/

Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden T.L. (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421. doi:10.1186/1471-2105-10-421.

Food and Agriculture Organization of the United Nations. (2001) Restoration of degraded forest land in Thailand:the case of Khao Kho. Unasylva 207:52-56.

Hernandez D., Tewhey R., Veyrieras J.B., Farinelli L., Osterås M., François P., Schrenzel J. (2014) De novo finished 2.8 Mbp Staphylococcus aureus genome assembly from 100 bp short and long range paired-end reads. Bioinformatics 30:40-49.

Kajitani R., Toshimoto K., Noguchi H., Toyoda A., Ogura Y., Okuno M., Yabana M., Harada M., Nagayasu E., Maruyama H., Kohara Y., Fujiyama A., Hayashi T., Itoh T. (2014) Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Research 24:1384-1395.

Li H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio.GN]

Lukmandaru G., Takahashi K. (2008) Variation in the natural termite resistance of teak (*Tectona grandis* Linn. fil.) wood as a function of tree age. Ann. For. Sci. 65:708-796.

Luo R., Liu B., Xie Y., Li Z., Huang W., Yuan J., He G., Chen Y., Pan Q., Liu Y., Tang J., Wu G., Zhang H., Shi Y., Liu Y., Yu C., Wang B., Lu Y., Han C., Cheung D.W., Yiu S.M., Peng S., Xiaoqian Z., Liu G., Liao X., Li Y., Yang H., Wang J., Lam T.W., Wang J. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1:18. doi: 10.1186/2047-217X-1-18.

MacCallum I., Przybylski D., Gnerre S., Burton J., Shlyakhter I., Gnirke A., Malek J., McKernan K., Ranade S., Shea T.P., Williams L., Young S., Nusbaum C., Jaffe D.B. (2009) ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads. Genome Biology, 10:R103 doi:10.1186/gb-2009-10-10-r103

Quillery E., Quenez O., Peterlongo P., Plantard O. (2014) Development of genomic resources for the tick *Ixodes ricinus*: isolation and characterization of single nucleotide polymorphisms. Mol Ecol Resour. 14:393-400.

Safonova Y., Bankevich A., Pevzner P.A. (2014) dipSPAdes: assembler for highly polymorphic diploid genomes. Research in Computational Molecular Biology 8394: 265-279.

Sahlin K., Vezzi F., Nystedt B., Lundeberg J., Arvestad L. (2014) BESST – Efficient scaffolding of large fragmented assemblies. BMC Bioinformatics 15:281 doi:10.1186/1471-2105-15-281

Stanke M., Diekhans M., Baertsch R., Haussler D. (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24: 637-644.

Zimin A., Marçais G., Puiu D., Roberts M., Salzberg S.L., Yorke J.A. (2013) The MaSuRCA genome Assembler. Bioinformatics 29:2669-2677 doi: 10.1093/bioinformatics/btt476.